METHOD AND APPARATUS FOR AUTOMATIC DATA ACQUISITION OF FORMS

BACKGROUND OF THE INVENTION

Cross Reference To Related Applications

[0001] Not applicable.

Statement Re Federally Sponsored Research

[0002] Not applicable.

Field of the Invention

[0003] The present invention refers to a method and an apparatus for automatic data acquisition of forms whose design and content are not known in advance, by inputting data and comparing the data with stored patterns.

Description of the Related Art

[0004] It is a problem for companies, organizations and others to organize information found in different types of paper forms, documents, and the like.

[0005] With new, modern technology, these items can be scanned with a scanner and entered into a database via commercially available software programs. However, sorting, identification and other checking routines must to a large extent still be performed manually via the computer's display or screen.

[0006] For example, to store an invoice from a company as one specifically designed document with a logotype and other visual elements, the invoice must be revised so that its format is adapted to one that can be accepted by the software and then stored in a database. This and other

procedures must be repeated each time an invoice with a new design is scanned and processed by the software.

[0007] To identify an invoice from a company that is already registered or known, the whole invoice is often searched, which is time consuming. Certain software programs have search routines that restrict the extent of this searching. It is, however, difficult to safeguard against blurred or hand-written lines of text, and the like, so restricted searches may prove inoperative.

[0008] A need therefore exists for all who handle invoices and other forms to quickly identify them and/or quickly enter and store new patterns. Prior inventions proved insufficient. U.S. Patent 4,933,979 described traditional data acquisition from forms and requires pre-defined templates or patterns with no self-learning (adaptive) ability. U. S. Patent 5,140,650 discloses data acquisition from forms with what is known as "Form Out" technology to cover-over the original document and only retain the parts that are "filled-in." This data acquisition is often combined with data acquisition according to U.S. Patent 4,933,979, but, as mentioned, this patent does not have any adaptive function for data acquisition of unknown forms. Another patent, U.S. Patent 5,293,429, discloses the classification of documents with the help of lines on the documents and does not directly concern data acquisition or any adaptive function. This patent does not ensure the identification of lines with object areas (areas with text) and a "RCG-value" (ReCoGnition, a number that uniquely identifies a document).

[0009] None of the above mentioned patents generates a form map for a form unknown to the system nor stores the map in real time in a form database for identification recognition at the next opportunity. The unknown forms must therefore be stored for later identification by other means.

BRIEF SUMMARY OF THE INVENTION

[0010] One of the objectives of the present invention is to solve the problems mentioned above during what is known as automatic data acquisition (interpretation) in connection with the handling of paper-based information.

[0011] The present invention concerns a method and an apparatus for the automatic data acquisition of forms where there is no prior knowledge of the forms' appearance or where on the forms information is to be found. In this way, templates of forms do not have to be defined in advance, but instead, the forms are registered as they are submitted in real time.

[0012] To accomplish the above objectives, the present invention specifies a method and an apparatus for the automatic data acquisition of forms whose design and content are not known in advance, by inputting the forms into a data processing unit together with the storage of form patterns. The method is adaptive; it includes learning and registering of forms as patterns without filled-in text, and it includes the following steps for accomplishing the adaptive registration: generation of a form map based on previously unknown form design for identifying information contained on the form, searching and comparing the form map with stored, registered maps, storing generated form maps in a storage means when a form map does not coincide with a stored map according to predetermined limits for agreement, providing an indication of agreement according to the limits for agreement when agreement is found, and continuing data acquisition for identifying informational content of the form. According to one embodiment of the present invention, the form map may include an object area list with objects contained in the form where the objects comprise color and/or text.

[0013] In an alternative embodiment, the form map constitutes a line map comprising objects in the form of colored lines from the form. Horizontal lines in the line map are used to produce a horizontal key by dividing the form into a predetermined number of horizontal segments along a y-axis in a cartographic system of coordinates, whereby each segment is equivalent to a position in the horizontal key. Vertical lines in the line map are used to produce a vertical key by dividing the form into a predetermined number of vertical segments along the x-axis in the cartographic system of coordinates, whereby each segment is equivalent to a position in the vertical key. At least one line element that is included in a segment is marked in the equivalent key position, and segments that lack line elements remain unmarked in the equivalent key position. A horizontal key and/or a vertical key constitute a line key in the line map, whereby during searching, the line key generated is compared with stored line keys for verifying agreement. The line keys are sorted in the storage means according to the number of markings.

[0014] The object's horizontal position in the object area list is used to generate a horizontal key by dividing the form into a predetermined number of horizontal segments along a y-axis in a cartographic system of coordinates, whereby each segment is equivalent to a position in the horizontal key. The object's vertical position in the object area list is used to produce a vertical key by dividing the form into a predetermined number of vertical segments along the x-axis in the cartographic system of coordinates whereby each segment is equivalent to a position in the vertical key. At least one object that is included in a segment is marked in the equivalent key position, and segments that lack objects remain unmarked in the equivalent key position. A horizontal key and/or a vertical key constitute an object key in the object area list, whereby during searching, the object key generated is compared with stored object keys for verifying agreement.

[0015] The object keys are preferably sorted in the storage means according to the number of markings. Searching results in a pre-defined number of requested probable candidates for the currently searched unknown forms. If needed, an operator can support manually the whole or parts of the adaptive registration or identification of the new form or registered forms, respectively, if several alternative candidates are found as probabilities according to a factor of merit. Finally, the identity of the form is confirmed by the data acquisition of a RCG-value.

[0016] Furthermore, the present invention specifies an apparatus for performing the above method. The apparatus carries out automated data acquisition of forms whose design and informational content is not known in advance, by inputting the forms into the data processing unit together with storage of form patterns. The device learns adaptively and registers the design of the form, and includes a computer system with the following means for carrying out the adaptive registration: means for generating a form map based on the previously unknown form design for identifying information contained on the form, means for searching and comparing the form map with stored, registered maps in a means for storing form maps, means for storing generated form maps in the storage means when they do not coincide with a stored map according to predetermined limits for agreement, means for indicating agreement according to predetermined limits for agreement when agreement is found, and means for identification and continued data acquisition of the informational content of the form.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWING

[0017] Further reference to the enclosed figures and associated text will give a clearer understanding of the present invention.

[0018] FIGURE 1 is a schematic view illustrating how a line pattern is accomplished from a scanned-in invoice.

[0019] FIGURE 2 is a schematic view of a flow-path for scanning, identifying, interpreting and validating an unknown form according to the present invention.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT OF THE INVENTION

[0020] In the continued description of the present invention, the forms are presented as invoices. However, the invention is not limited to invoice forms but also covers general documents containing text, figures, and the like, as forms. Invoices are used here as an example of a form to exemplify the invention.

[0021] Referring now to FIG. 1, there is illustrated schematically one part of an invoice 10 that is scanned into a computer and that is shown on a display. As is evident from the invoice 10, it is unclear or blurred after the scanning or input. The invoice consists partly of a logotype 12, a vertical line element 14 and a horizontal line 15 element. Note that even the logotype contains long black lines or varying degrees of shaded colored line elements 16 that have been partly registered in a line map 18 according to the present invention. The line elements give an idea of what the original logotype 12 looked like, which simplifies identification when the invoice is an object to be identified as being registered in a form map database. Colored lines also include grey scales. The form map, in this case the line map 18, has been filtered from other objects 19, such as text objects or colored objects. Even the line elements that include color, which cannot be reproduced here, can be included as many colored fields on a form, such as the invoice 10.

[0022] The invoice 10 that is prepared according to the present invention, hereafter being designated EH (Eyes & Hands), must be identified at an early stage. For successful

identification, EH must have on a previous occasion, learned what the current invoice 10 looks like, which in simple terms means that information about the invoice is available in the form database in EH. By necessity, the identification must be quick and be able to be made in a database that holds a very large number of invoice identities or line maps 18. It is not uncommon for databases to contain more than ten thousand identities 18.

[0023] The method and apparatus that EH uses does not require that an invoice always be put through a scanner in exactly the same way, i.e., the information on the invoice can vary somewhat in the x and y axes within a predetermined measurement or threshold value. FIG. 1 includes a schematic cartographic system of coordinates. In the present invention, EH searches for all vertical and horizontal line elements 14, 15 of a predetermined length on the invoice. Line elements 14, 15 do not need to be free-standing and isolated, but can, for example, be part of a larger logotype text 12, such as "ReadSoft AB" shown in FIG. 1. The logotype 12 is represented as the line element 16 in the line map 18.

[0024] The horizontal and the vertical line elements constitute the basis for generating a horizontal key (h-key) and a vertical key (v-key), respectively, according to the following. The invoice is divided into a large number of horizontal segments along the y-axis (not shown). Each segment is equivalent to one position in the h-key. If a certain segment includes one or more horizontal line elements 15, a mark or tag is placed in the equivalent key position. If not, an empty space, an inverted mark or anything else that differentiates itself from a mark, is used instead. A v-key for the vertical line elements 14 is generated in a similar manner along the x-axis. The v-keys are given designations and together with the h-keys constitute a line key. Following this, a search is performed, which means that the current line key is compared with line keys for known invoices that exist in the EH database. This comparison takes into account

that individual line elements 14, 15 can vary somewhat in position, and that the total pattern of lines can be displaced somewhat according to suitable predetermined values in the x and y directions, horizontally and vertically, respectively. The line keys in the database are sorted according to the number of markings (tags), which are used to make the searching effective. The search results in a predetermined number of probable candidates for the identity of the current invoice. All candidates are associated with a factor of merit or a probability that they are the current invoices. The identity of the invoice is finally confirmed by carrying out an interpretation of what is known as the RCG-value (RCG- ReCoGnition). The RCG-value is a value at a given position that is unique for a certain invoice or other form. Examples of such values are bank giro numbers, post-office giro numbers, invoice numbers, total amounts, and the like.

[0025] The segments generated can, for example, form checked patterns that are fine-screened to varying extents according to the relative need for rapid searching. The line keys can even be implemented on objects formed wholly or partly of text and colors. These are assigned line keys from an object area list that includes x and y-keys for the object. The object area list can, for example, include positions for certain selected objects. The principles for line maps stated above are even appropriate for objects other than line elements to accomplish identification of forms. If the line keys are not found in the database, there is an indication that the invoice is not known which results in new line keys being stored in the database so that the database is updated in real time. If necessary, an operator can, via his/her computer, manually support, in whole or part, the adaptive registration and/or identification of a new form or registered form, respectively, if several alternative candidates are presented as probable according to given factors of merit.

[0026] In addition, the present invention includes an apparatus for performing the method according to the above disclosure. The apparatus performs the automatic data acquisition of forms whose design and informational content are not known in advance, by inputting the forms together with the storage of form patterns. The apparatus registers in an adaptive manner and learns the design of forms, and includes a computer with the following means for accomplishing the adaptive registration. There are means for generating a form map based on the previously unknown form design and for identifying information contained on the form. There are means for searching and comparing the form map with stored, recognized maps in a means for storing form maps. There are means for storage of generated form maps in the storage means when they do not coincide with a stored map according to predetermined limits for agreement. There are means for indicating agreement according to the limits for agreement when agreement is found, and there are means for identification and continued data acquisition of information contained on the form.

[0027] The means identified above are preferably controlled by computer hardware and software, such as, for example, a scanner for acquisition of data, an electronic storage medium (hard disk, CD-ROM, or the like) for storing information, signs, icons, signal generators, etc. for indicating purposes, and filters and comparitors so that searching and comparing may proceed, as well as filters and registers for identification purposes.

[0028] On the whole, the means used in the present invention are well known to a skilled person in the technical field, but the way in which they are coordinated to achieve the objects of the invention are innovative.

[0029] In one embodiment of the present invention, with reference to FIG. 2, a schematic flow-path is illustrated to show the scanning, identification, interpretation, and validation of a form.

[0030] FIG. 2 is divided by lines into areas to clarify the different steps in the method according to the invention, whereby the steps include the scanning of the form 200, identifying the form 210, interpreting the form 220, plus validating the form 230.

[0031] The form is scanned 200 into EH, and identification 210 follows. Identification consists of generating a line map 212, or alternatively, an object area list, whereby a line key is generated. Following this, the form is compared 214 with known keys in the form map database, whereby a conformation of identification is obtained via the RCG-value. The next step includes deciding whether the identification was successful 216 according to the conditions "Yes" or "No". If the decision results in "No", a conditional investigation is made to determine if there are more candidates in the form of line keys 218. If the answer here is "Yes", a loop is performed until a successful identification is finally made, or until no further line key candidates are presented.

[0032] In the case of a successful identification, interpretation 220 of the form then begins by interpreting with the help of the current or existing form maps 222, after which validation 230 or evaluation 232 of the fields of the form takes place. As an option, the operator can assist with selection if several alternative fields are found 234.

[0033] If the identification 210 is unsuccessful, and no further line keys are presented 218, interpretation 220 is performed in that self-learning with a form definition 224 is accomplished. The form definition consists of a template or a set of rules that describes the common elements of a specific collection of forms, for example, Swedish invoices. Following this, the RCG-value

is interpreted 226 and a decision is made 228 whether the current RCG-value can be found in the form database. If the answer is "Yes", a re-interpretation begins 229, followed by a continued interpretation 222 that leads to validation 232.

[0034] If, on the other hand, the answer is "No", validation commences 230, 236, after which the form is saved in the form map database with the line key 238. Prior to steps 236 and 238, the operator can, if several field alternatives are found, assist with the self-learning process.

[0035] The embodiments of the present invention described above are possible embodiments, but are not intended to limit the invention to such, as further embodiments will be evident to a skilled person in the technical area..